REMARKS

Applicants gratefully acknowledge Examiner Vicary for taking time from his busy schedule and for courtesies extended during a telephone interview dated July 15, 2008, including co-inventors Dr. Gustavson and Dr. Gunnels.

During this interview, the Examiner indicated his concern about using "non-standard format" in the claims without more completely defining the term and clarifying which specific non-standard format is being claimed. The Examiner also agreed that Applicants would be able to incorporate the description from any of the copending applications into the specification of the present application, in order to more specifically define the terms considered by the Examiner as desirable to add to the claims.

Applicants believe that the above claim and specification amendments and specification should appropriately address the Examiner's concerns and request that the Examiner contact Applicants' representative at the number below if additional changes are deemed necessary for immediate allowance.

Claims 1-9 and 11-19 are all the claims presently pending in the application. Claims 10 and 20 are canceled.

It is noted that Applicants specifically state that no amendment to any claim herein, if any, should be construed as a disclaimer of any interest in or right to an equivalent of any element or feature of the amended claim.

In accordance with the above-mentioned telephone interview, the independent claims have been amended:

- to clarify the non-standard format as including the register block format mentioned in line 13 of page 7 and lines 8-12 of page 16 of the originally-filed specification;
- to clarify the three data streams, as discussed at line 21 of page 14 through line 2 of page 15 of the originally-filed specification; and
- to clarify the multiple loading as reference to SIMD capability.

Also in accordance with the above-mentioned telephone interview, the specification has been amended to provide more information from co-pending application S/N 10/671,888 that more fully defines the register block format. The seven co-pending applications including the present application provides three examples of new data structures using non-standard format (e.g., the register block format, composite blocking, and the hybrid full

packed format. The prefetching of the present application could be using any combination of these three non-standard formats, although the inventors consider the register blocking format to be particularly relevant to the present invention because it is predetermined to be based on loading into the register set of the FPUs.

Claims 1-9 and 11-19 stand rejected under non-statutory double patenting over claims 1, 3-6, 8-12, and 14-19 of co-pending application S/N 10/671,937, further in view of "Superscalar GEMM-based Level 3 BLAS" to co-inventor Gustavson et al.

Claims 1-9 and 11-19 stand rejected under 35 U.S.C. § 112, first paragraph, as failing to comply with the written description requirement.

Claims 1-9 and 11-19 stand rejected under 35 U.S.C. § 112, second paragraph, as being indefinite.

Claims 1-9 and 11-19 stand rejected under 35 U.S.C. § 102(b) as anticipated by Gustavson et al., "Superscalar GEMM-based Level 3 BLAS – The On-going Evolution of a Portable and High-Performance Library."

These rejections are respectfully traversed in the following discussion.

I. THE CLAIMED INVENTION

The claimed invention is directed to a method of executing a linear algebra subroutine on a machine having at least one floating point unit (FPU) with one or more associated load/store units (LSU) to load data into and out of floating point registers (FRegs) of the FPU by way of an L1 cache.

For an execution code controlling an operation of said floating point unit (FPU) performing a linear algebra subroutine execution, instructions are inserted to move data into a cache providing data for the FPU so that the LSUs can move the data into the FRegs before it is scheduled to by used by said linear algebra subroutine execution. The data being prefetched into the cache from memory is in a nonstandard format predetermined to reduce a number of data streams for a level 3 linear algebra processing to be three streams and to allow a multiple loading of these streams into said FPU by said LSU,

The nonstandard format comprises a register block format wherein data is stored in blocks of size p-by-q, where p and q are small integers so that the pieces of these blocks can be fitted into the FRegs. The three data streams comprise data of one matrix of the level 3

linear algebra processing as considered to be resident in the cache and the two remaining matrix operands of the level 3 linear algebra processing as residing in a cache level higher than the cache.

Conventional compilers do not have the capability to automatically pre-fetch (timely move) data into the FPU for Level 3 Dense Linear Algebra Subroutines, particularly in view of the newer architectures having FPUs and LSUs.

The claimed invention, on the other hand, teaches how to timely load data into cache, using a non-standard format predetermined to allow the minimum of three data streams and to allow multiple loading into the FPU. This feature can also be accomplished by conventional compilers, when modified to incorporate the concepts of the present invention.

II. THE DOUBLE PATENT REJECTION

The Examiner continues to consider that all pending claims of the present application are obvious over the claims of co-pending application 10/671,937.

As indicated during the above-mentioned telephone interview, the seven co-pending applications listed at the beginning of the present application (including the present application) are considered as capable of working together to improve efficiency on the newer SIMD machines in a synergistic manner. As also indicated during the telephone interview, all seven of these co-pending applications were filed on the same day so that there would be no difference in patent term, excluding possible term extension differences, thereby rendering a terminal disclaimer as moot.

More important, as indicated during the telephone interview, Applicants consider these seven applications as patentably distinct improvements in the art. As such, each of the seven applications has independent claims specifically intended to cover their perceived novelty of that application alone, even if various dependent claims might cover subject matter of another co-pending application, given the potential for synergy with another technique. Thus, the various claim sets of these seven co-pending applications clearly cover distinctly different scopes, when viewed from the perspective of the different independent claims.

Of particular relevance to the alleged double patenting of the rejection currently of record, Applicants bring to the Examiner's attention that the preloading described in co-pending application 10/671,937 is actually an alternative to the prefetching method of the

present invention. That is, the prefetching of the present invention uses the nonstandard format of register blocking, which has placed the matrix data into a contiguous format of blocks of data designed for loading onto the FPUs. Therefore, the prefetching of the present invention does not subsequently utilize the technique of preloading described in 10/671,937. Rather, the preloading of 10/671,937 is an alternate method to overcome a one or more cycle (a 5-cycle penalty was discussed in the disclosure for an older machine) penalty associated with the cache/FPU loading of the newer machines. The FPU has associated loading instructions. Thus, incorrectly loaded data can be re-arranged inside the FPU to be in a correct format. See the description in co-pending application YOR920030169 (US Patent S/N 10/671,888) where we describe two errors correcting each other via the pseudo matrix concept.

The present invention achieves the same effect as the preloading described in 10/671,937 because of the use of the register block format in the prefetching. Re-arrangement of data is not required in the FPU as it is already in the correct format for FPU processing.

Thus, contrary to the Examiner's position, the prefetching and preloading described respectively in the present application and in copending application 10/671,937 are not sequential processes as described in the rejection currently of record. Rather, they are alternate processes and, therefore, clearly patentably distinct.

Turning now to the first full paragraph on page 5 of the present rejection, the Examiner alleges that the Gustavson article in §3.1 describes a non-standard format for prefetching:

"*... Gustavson discloses, said data being prefetched into cache from a memory in a nonstandard format (section 3.1, first indented paragraph of page 210, technique of keeping a small square block of C in registers; this technique of prefetching C in the format of a small square block as opposed to the prefetching o f A and B can be considered to nonstandard)*"

In response, Applicants bring to the Examiner's attention that the description in §3.1 does not make any suggestion whatsoever about the format used for prefetching, since it merely describes keeping a small block of C data in the registers, which is an entirely different concept from that of describing the format used to prefetch the small block of C data kept in these registers. There is no suggestion in this article to use a non-standard format for prefetching data into the cache and the Examiner's interpretation of this section in Gustavson

is clear evidence of improper hindsight, since the Examiner is clearly attempting to extend the activity within the FPU registers as indicative of the format used to prefetch data from memory into cache.

There is no correspondence whatsoever in the prior art between the format of data in the FPU register set and the retrieval of data from memory. As Applicants keep pointing out, the conventional storage/retrieval of data is based entirely upon the standard for each computer language, either row major or column major.

Relative to the Examiner's allegation that Gustavson's article described three streams for the level 3 processing, Applicants explained during the above-mentioned telephone interview that pre-SIMD machines were not capable of reducing the data streams down to only three streams. Rather, for each of the three operands, there were pluralities of streams dependent upon the size of the data blocks for each operand.

In paragraph 43 on page 17 of the Office Action the Examiner concedes that Applicants that prefetching data into cache would be different from preloading data from cache into the FPU registers but insists in the final sentence that "*... pre-loading can nevertheless necessitate pre-fetching as well (which does not mean that they are not the same).*"

In response, Applicants again point out that prefetching and preloading are alternative methods that solve a one or more cycle (5-cycle was used in the Spec) penalty at the cache/FPU interface of newer architectures.

Therefore, contrary to the Examiner's rationale in the rejection of record, Applicants respectfully submit that the preloading of matrix data in the Gustavson article, as presuming that a prefetching of that data precedes the preloading, would not render the present prefetching obvious over the co-pending application related to preloading since the machine in the Gustavson article used prefetching and preloading in a different context so that they were sequential operations.

In view of the above, Applicants again submit that the preloading technique of co-pending application S/N 10/671,937 (e.g., YOR920030171US1) does not render obvious the prefetching technique of the present application S/N 10/671,889 (e.g., YOR920030170US1), further in view of the prefetching/preloading techniques described in Gustavson's previous article.

Therefore, the Examiner is again respectfully requested to reconsider and withdraw

this rejection.

III. THE 35 USC §112, FIRST PARAGRAPH, REJECTIONS

Claims 1-9 and 11-20 stand rejected under 35 U. S.C. §112, first paragraph, as allegedly failing the written description requirement. Applicants believe that the above claim amendments, consistent with Applicants' understanding during the above-mentioned telephone interview, and the above specification amendments incorporating additional description from co-pending application S/N 10/671,888 (IBM docket YOR920030169US1) on the register block format and from co-pending S/N 10/671,935 (IBM docket YOR920030330US1) on six level 3 L1 kernel routines, appropriately address some of the concerns in this rejection.

In view of the above, Applicants respectfully request that the Examiner reconsider and withdraw this rejection.

IV. THE 35 USC §112, SECOND PARAGRAPH, REJECTION

The Examiner objects to various terms in the claims. In view of the discussion in the above-mentioned telephone interview, Applicants believe the above claim amendments appropriately address the Examiner's concerns.

In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

V. THE PRIOR ART REJECTION

The Examiner alleges that the article "Superscalar GEMM-based Level 3 BLAS –The On-going Evolution of a Portable and High-Performance Library," Para '98, pages 207-215), co-inventor Gustavson, et al., teaches the claimed invention.

In response, Applicants submit that this paper refers only to multiple loads of load multiple type k=1. The present application addresses architectures capable of a SIMD load with $k > 1$, and this paper by Gustavson et al., does not cover the specific situation that the present invention can address for the newer architectures.

Moreover, the independent claims of the present application also refer to non-standard format used when pre-fetching data from memory into L1 cache. This aspect of independent claim 1 is not present in this paper. In paragraph 3 on page 12 of the Office Action, the Examiner relies upon the description in section 3.1, first indented paragraph of page 210 of this paper, presumably meaning the following:

" *This technique, to keep a small square block of C in registers and replace entries of A and B between consecutive iterations of the innermost loop, maximizes the ratio between the number of MAAs and the number of load and store instructions, used to transfer data to and from registers, i.e., #MAAs/(#LOADs + #STOREs) is maximized.*"

However, Applicants respectfully disagree with the Examiner that one having ordinary skill in the art would agree with the Examiner's characterization that this description has anything at all to do with data format, let alone data format used for data moved into L1 cache. Indeed, Applicants respectfully submit that these words have no suggestion whatsoever as to whether the data anywhere in this sentence is in any specific format, since the mechanism is not described as dependent upon any such specific data format.

Therefore, Applicants submit that there is no suggestion in co-inventor Gustavson's cited paper concerning non-standard format for data transfer between memory and cache, as required by the plain meaning of the claim language.

Hence, turning to the clear language of the claims, in Gustavson there is no teaching or suggestion of: " ... for an execution code controlling an operation of said floating point unit (FPU) performing a linear algebra subroutine execution, inserting instructions to move data in a contiguous and stride one format either into a cache providing data for said FPU for direct loading into said FPU, so that said LSUs can load said data into said FRegs before it is scheduled to by used in said linear algebra subroutine execution, said data being prefetched into said cache from a memory in a register block format predetermined to reduce a number of data streams for a level 3 nested loop matrix-matrix type kernel type operation processing to be three streams and to allow a loading of these streams into said FPU by said LSU, said register block format comprising a data storage format wherein data is stored in blocks of size p-by-q where p and q are small integers so that the pieces of these blocks can be fitted into said FRegs, and wherein said three data streams comprise data of one matrix of said level 3 processing as considered to be resident in said cache and data for two remaining matrix

operands of said level 3 processing as residing in a cache level higher than said cache", as required by independent claim 1. The remaining independent claims have similar language.

In view of the above, the Examiner is respectfully requested to withdraw this rejection.

VI. FORMAL MATTERS AND CONCLUSION

In view of the foregoing, Applicant submits that claims 1-9 and 11-19, all the claims presently pending in the application, are patentably distinct over the prior art of record and are in condition for allowance. The Examiner is respectfully requested to pass the above application to issue at the earliest possible time.

Should the Examiner find the application to be other than in condition for allowance, the Examiner is requested to contact the undersigned at the local telephone number listed below to discuss any other changes deemed necessary in a telephonic or personal interview.

The Commissioner is hereby authorized to charge any deficiency in fees or to credit any overpayment in fees to Assignee's Deposit Account No. 50-0510.

Respectfully Submitted,

Date: July 30, 2008

Frederick E. Cooperrider
Registration No. 36,769

McGinn Intellectual Property Law Group, PLLC
8321 Old Courthouse Road, Suite 200
Vienna, VA 22182-3817
(703) 761-4100
Customer No. 21254

CERTIFICATION OF TRANSMISSION

I certify that I transmitted electronically, via EFS, this Amendment under 37 CFR §1.116 to the USPTO on July 30, 2008.

Frederick E. Cooperrider (Reg. No. 36,769)